



University of Groningen

How reliable are radiocarbon laboratories? A report on the Fourth International Radiocarbon Inter-comparison (FIRI) (1998-2001)

Boaretto, E; Bryant, C; Carmi, [No Value]; Cook, G; Gulliksen, S; Harkness, D; Heinemeier, J; McClure, J; McGee, E; Naysmith, P

Published in:
Antiquity

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2003

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Boaretto, E., Bryant, C., Carmi, N. V., Cook, G., Gulliksen, S., Harkness, D., ... Van Styrdonck, M. (2003). How reliable are radiocarbon laboratories? A report on the Fourth International Radiocarbon Inter-comparison (FIRI) (1998-2001). *Antiquity*, 77(295), 146-154.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

How reliable are radiocarbon laboratories?

A report on the Fourth International Radiocarbon Inter-comparison (FIRI) (1998–2001)

Elisabetta Boaretto², Charlotte Bryant¹, Israel Carmi², Gordon Cook³, Steinar Gulliksen⁴, Doug Harkness⁸, Jan Heinemeier⁵, John McClure⁸, Edward McGee⁶, Philip Naysmith³, Goran Possnert⁷, Marian Scott^{8*}, Hans van der Plicht⁹, Mark van Strydonck¹⁰.

Radiocarbon laboratories undertake rigorous programmes of internal quality control (QC) and overall quality assurance (QA). In a laboratory “inter-comparison”¹ samples of the same age are dated at different laboratories using a range of techniques and the results are then compared. The authors summarise the results of the fourth of these scientific audits.

Keywords: Radiocarbon dating, quality assurance, outlier, laboratory inter-comparison

Rationale

The most recent radiocarbon inter-comparison exercise (FIRI), completed in 2001, was also the most extensive so far, with 85 laboratories participating. The study was designed firstly to assess the comparability of the results from different laboratories and then to quantify the extent and possible causes of any inter-laboratory variation. Radiocarbon dating is universally employed as a dating tool in archaeology, but there is an inevitable diversity of experimental approaches within radiocarbon dating facilities and in this situation the issue of comparability of results amongst laboratories becomes paramount. In keeping with the principles of analytical science, radiocarbon laboratories have always been conscious of the importance of accuracy and precision for their reported results i.e. the ethos of analytical quality control (QC) which in turn is the foundation for the wider concept of quality assurance (QA). The care and effort given to establishing and maintaining primary standards and reference materials exemplify this concern for good quality management within the radiocarbon community.

As early as 1989, Long and Kalin (1990) stressed that it was incumbent on individual radiocarbon laboratories to engage in a formal programme of quality assurance (QA) while

1) NERC Radiocarbon Laboratory, Scotland, 2) Weizmann Institute, Israel, 3) SUERC, Scotland, 4) NUST, Norway, 5) University of Aarhus, Denmark, 6) University College Dublin, Eire, 7) University of Uppsala, Sweden, 8) University of Glasgow, Scotland, 9) University of Groningen, The Netherlands, 10) KIK, Belgium.

* corresponding author: marian@stats.gla.ac.uk

Polach (1989) noted that the opportunity for internal checking by individual laboratories in routine ^{14}C measurement was hampered by a lack of suitable quality control (QC) and reference materials. The work reported here describes ongoing international efforts by means of a laboratory inter-comparison to assure users of laboratory quality and comparability of measurements and to provide suitable quality control and reference materials. This work builds on the previous laboratory inter-comparisons that have taken place over the last 20 years (ISG, 1982; Scott et al, 1991; Rozanski et al, 1992; Gulliksen & Scott, 1995).

A substantial effort has been made by the ^{14}C community to develop and apply both internal and external QA procedures. FIRI provides a part of these procedures in the form of an independent check of laboratory performance. However, it only provides a spot check of operational performance at the time it was carried out and does not measure consistent performance over a period of time and so should not form the basis of a 'league table of laboratory performance'. This is why the FIRI results are published without laboratory attribution.

Objectives

The specific objectives of the Fourth International Radiocarbon Inter-comparison (FIRI) (Bryant et al, 2000) were to provide:

- An unambiguous demonstration of the degree of consistency or otherwise among the results obtained, on a routine basis, from different laboratories. This information is crucial for both laboratories and procurers (researchers and funding agencies).
- Quantification of the extent of, and identification of, the possible causes of, any inter-laboratory variation.
- Direct assessment of the comparability of liquid scintillation counting (LSC), gas proportional counting (GPC) and accelerator mass spectrometry (AMS) techniques.
- Creation of suitable, well-characterized quality control and reference materials.
- Assurance of trace-ability of the measurements and provision of an independent check on laboratory performance.

These objectives are directly related to analytical quality control by focusing on experimental accuracy, precision and reproducibility as indices for the assessment and inter-comparison of laboratory performance. Evidence of an acceptable level of analytical quality control is the essential precursor for overall quality assurance. The basic methodology employed in the inter-comparison was to invite laboratories to date a series of reference samples so as to compare their performances and the performance of different radiocarbon methods.

Selection, preparation and testing of control samples

The selection and preparation of large samples of homogeneous ^{14}C activity (uniform age) which can then be sub-divided are vital components of a successful radiocarbon inter-comparison exercise. Natural materials were sought which were representative of routinely dated materials and whose ages spanned the full range of the applied ^{14}C timescale. Potential materials that were identified included wood (if possible with a dendrochronological date), peat, bone, marine carbonate and grain, together with specific components of samples such as the cellulose fraction of wood and the humic acid fraction of peat. The degree of preparation varied from a thorough physical mixing (*e.g.* marine carbonate – turbidite sediment), through

grinding and mixing (whole peat), to complete chemical homogenisation (humic acid extraction from peat).

All bulk materials were prepared in a single batch, homogenised and checked by replicate analyses on eight randomly selected aliquots.

Homogeneity testing

Other than for the dendrochronologically dated wood samples, the bulk samples were tested at different sub-sample sizes (reflecting one of the key differences between AMS and radiometric measurement). In all cases, two laboratories checked the sample homogeneity. All the samples were in good agreement with the exception of the turbidite and the modern cellulose samples. For the turbidite, the difference between the two laboratories was later demonstrated to be due to pretreatment. For the modern cellulose sample, the difference between the two laboratories was due to a small error in assessing the modern standard activity on the part of one laboratory. Notwithstanding these difficulties, the results of the homogeneity testing indicated that when laboratories complied with specific instructions concerning sample handling and pre-treatment, all of the samples could be considered to be homogeneous and thus suitable for inter-comparison.

Tests conducted by the participating laboratories

Each laboratory participating in the inter-comparison was invited to measure a total of ten samples drawn from a set of seven core materials within a one-year period. These samples are described in Table 1. They included four dendro-dated samples from the Belfast and German master chronologies, to provide an assessment of laboratory accuracy. Three sets of duplicate samples were provided blind (Kauri wood, Belfast dendro-dated wood and Barley mash) to allow an assessment of laboratory precision.

The reference samples were distributed to over 120 laboratories during 1999 and by the deadline of December 2000, 92 sets of results had been received, with some laboratories submitting more than one set of results. The broad geographical distribution of participating laboratories is shown in Table 2, the radiocarbon techniques employed in Table 3 and a list of the participating laboratories can be found in Table 6.

Table 1. Core sample descriptions

Core Sample description	FIRI code*	Age/Activity
Kauri wood	A, B	Near background
Marine turbidite	C	~3 half-lives
Belfast dendro-dated wood	D, F	~1 half-life
Humic acid	E	~2 half-lives
Barley mash	G, J	modern
Hohenheim dendro-dated wood	H	< 1 half-life
Belfast dendro-dated cellulose	I	~1 half-life

* Code to indicate the material which is being dated

Table 2. Geographical distribution of participating laboratories

Geographical area	Number of laboratories
Europe (EU)	35
Europe (non EU)	15
North and South America and Canada	16
Asia and the Far East	15
Australia and New Zealand	4

Table 3. Methods applied

Method	Number of laboratories using it
LSC ¹	44
GPC ²	19
AMS ³	17
Target feeder for AMS ⁴	8
Direct absorption and LSC ⁵	2

¹Liquid scintillation counting. ²Gas proportional counting. ³Accelerator mass spectrometry. ⁴Laboratories that prepare samples and send them to AMS laboratories for measurement. ⁵Laboratories that absorb sample carbon in the form of CO₂ into a tertiary amine or similar compound and measure the activity by LSC (generally a low precision method)

Relative performance of laboratories

A total of 122 observations out of 1056 (*i.e.* slightly over 10%) was identified as anomalous (*i.e.* outliers). From the statistical definition of an outlier, their proportion should have been around 5%. Thus the number of outliers was approximately twice the number that would be expected if they were occurring purely by chance. 39 laboratories (42%) had at least one result classed as an outlier. Of the 39, almost 60% (23) of these had more than one of their results thus classed and over 20% (9) had five or more such results. The distribution of outliers was uniform over the 10 samples, thus, no single sample-type contributed the majority of the outliers. Of the 122 outliers, 87% came from LSC laboratories.

Other sources of variation (pre-treatment, modern standard and background material)

For the turbidite sample, a significant age difference was observed between the acid-leached and the non pre-treated samples. For the whole wood samples, a very small – so practically unimportant, but statistically significant – effect due to pre-treatment was also observed. There was an indication of an association between the presence of an outlier and the modern standard used by these laboratories: further analysis indicated that the presence of outliers was linked to the modern standard used, some laboratories having no access to the primary standards of NIST OxI and OxII. After omission of outliers, there was then no evidence of a difference, on average, for any sample due to modern standard or background materials, with the exception of the near background sample (Kauri wood).

Overall, a relatively small number of laboratories (14%) generated more than 60% of the outlying observations, and the majority of these laboratories use liquid scintillation techniques (including direct absorption). However, it should be noted that there remain a substantial number of liquid scintillation laboratories with none or only one outlier.

Measures of precision and accuracy – comparing duplicates

Laboratories were asked to measure three pairs of duplicate samples: A and B (Kauri wood, near background activity), D and F (Belfast wood, around 50 pMC (percent modern carbon)) and G and J (barley mash, at approximately 111 pMC) to allow the assessment of laboratory precision relative to the quoted errors. The summary statistics for the differences of the duplicates are shown in Table 4 (note that D/F results are given in years BP).

This analysis showed that, on average, the difference between duplicates is zero (over all laboratories and also for individual laboratories). However, the magnitude of the difference in some individual cases was large relative to the quoted errors (and larger than expected given the interpretation of the quoted error). The implication is that, in such cases, a source of variation may not be completely accounted for in the quoted error. On the other hand, evidence was also observed of agreement between the duplicates, which was in fact better than would be expected on the basis of the quoted errors. This corresponds to an under-estimation of precision. The observed differences were adequately described by the quoted errors for approximately 50% of the laboratories.

Samples of known age

Accuracy can only be assessed against known age materials and for ^{14}C these are typically dendro-dated wood samples, so four such samples were included in FIRI. Consensus values (based on an iterative procedure involving the calculation of a weighted average (Rozanski *et al.*, 1992)) for the samples are shown in Table 5. A different method, based on reliability analysis, was used for the calculation of the consensus value for samples A and B.

Table 4. Descriptive statistics: (Differences between duplicates DF in years BP)

Sample pair	Number of results	Average or Mean difference	Standard Deviation of difference	Minimum difference	Maximum difference
AB	54	0.029 pMC	0.214	-0.66	0.53
GJ	71	-0.094 pMC	1.085	-4.37	2.76
DF	79	17.4 years BP	97.3	-239	310

Table 5. Consensus values

Sample	Known age	Consensus value (estimated 1 precision)
AB (pMC)	–	0.24 pMC ¹ (95% CI (0.23 – 0.30))
C (yBP)	–	18176(10.5) yBP ²
DF (yBP)	3200–3239BC (^{14}C age 4495BP)	4508 (3) yBP
E (yBP)	–	11780 (7) yBP
GJ(pMC)	–	110.7 (0.04) pMC
H(yBP)	313–294BC (^{14}C age 2215BP)	2232(5) yBP
I (yBP)	3299–3257BC (^{14}C age 4471BP)	4485(5) yBP

¹ percent modern carbon; ² radiocarbon years before present where present is 1950.

The four dendro-dated wood samples included in the list of core samples were D and F (duplicates) from the Belfast master chronology and dendro-dated to 3200–3239 BC (^{14}C age of 4495 BP); sample I (also from the Belfast master chronology) which has a dendro-date of 3299–3257 BC (^{14}C age of 4471 BP) and sample H from the German oak chronology which was dendro-dated to 313–294 BC (^{14}C age of 2215 BP). With respect to the dendro-dated samples, it can be observed that the consensus values and the average 'master' values are such that the differences are all within the limits of the quoted errors. Thus, the consensus results are in agreement with the master chronology results, so that overall, we can conclude that laboratories are, in general, accurate. For an individual laboratory, the difference (*known age-laboratory measured age*) for the dendro-dated samples can also be used to assess accuracy. It was found that the differences were distributed around zero, with the majority of results lying in the range of 100 years. Formal calculations showed that approximately 30% of the laboratories had a statistically significant offset.

Improving quality

The reported results from FIRI for each laboratory are in some senses a summary and therefore do not allow further examination of the causes of laboratory offsets (beyond that already reported here). The responsibility for investigating sources of the offset (and if required, amending procedures) rests with the individual laboratories. We have studied the effect of the modern standards and the background materials that are used by laboratories in their analyses and find no evidence that these factors make a significant contribution to the overall variation observed.

In-house policies for the definition and use of standards is important. The FIRI results demonstrate that there remains a need for standards and reference materials to which laboratories have ready access to allow checking and correction. Five categories of standard can be defined for application in the ideal situation viz.,

- i) Primary (or modern) standard. The internationally calibrated and certified materials NBS–Ox I and NBS–Ox II.
- ii) Secondary standards. Internationally recognised materials such as ANU-sucrose, Chinese – sucrose and the IAEA quality control reference series (C1–C8). Reference materials from TIRI and FIRI are now also available to expand this list.
- iii) In-house/ inter-laboratory QC standards. Materials selected to represent the type and age of submitted samples.
- iv) In-house working standard(s). A bulk supply of homogeneous material that is available in sufficient quantity to allow repeated and frequent analysis. These measurements are intended to monitor and control the reproducibility of the analytical process over time.
- v) Background standards. To achieve accurate and reproducible work, and especially with samples older than say four half-lives, it is essential to define the appropriate background signal using " ^{14}C free" (geologically old) material that has a chemical composition close to that of the sample. The background material should also be subjected to an identical form of any pre-treatment that is applied to the raw sample.

It is clear that programmes such as FIRI are, and will continue to be, necessary. One plan under consideration is that a major inter-comparison, such as FIRI, would be organised every four years but that in each of the three preceding years, a small number of samples

would be sent to laboratories to be analysed in a short time and feedback then given. In this way, the 'spot-check' nature of FIRI and the lack of continuous monitoring of performance would be remedied. Such a system would have benefits to the participating laboratories and would also provide a better 'quality guarantee' to the user communities. All results and a full report on the inter-comparison will appear as a special issue of *Radiocarbon* in 2003 (Scott *et al.*, forthcoming).

Rewards for the archaeological user

The selection by a user of a laboratory to which samples are sent is dependent on a number of factors, including cost and the time taken to obtain results. The choice may also be dependent on the sample material, on the sample size and on the precision with which the result is required. Laboratories differ in their capabilities to measure very small samples (AMS rather than radiometric); a few laboratories are able to measure to extremely high precision (<15 years) (only radiometric), while specialised pre-treatment procedures for unusual materials may only be available in a small number of laboratories.

But laboratory selection should also be based on an evaluation of the QA performance of the individual laboratory. By the simple fact of participation in a programme such as FIRI, a laboratory is emphasising its commitment to the quality assurance of its results. As in previous inter-comparisons, although the laboratory attribution in any results table is anonymous, a list of the participating laboratories is published (Table 6). Further, individual laboratories are encouraged, if they wish, to publish and publicise their own performance in the inter-comparison and many do so. Users are also encouraged to ask laboratories about their QA policies and should pay attention to laboratory participation in inter-comparisons and more specifically to laboratory use of standards and reference materials (such as TIRI, FIRI and IAEA C1–C8).

Table 6. Laboratories participating in FIRI

Laboratory name	Country	Laboratory name	Country
LATYR, La Plata	Argentina	Kyushu Environmental Evaluation Association	Japan
Pabellon INGEIS	Argentina	Institute for Advanced Science, Osaka	Japan
CSIRO, Glen Osmond	Australia	Palynosurvey Co	Japan
ANTARES AMS Centre, ANSTO	Australia	CCR Nagoya University	Japan
Arsenal Research	Austria	Gakushuin University, Tokyo	Japan
VERA, Universitat Wien	Austria	Kyoto Sangyo University	Japan
VRI, Institut für Radiumforschung und Kernphysik	Austria	Seoul National University	Korea
IRPA, KIK	Belgium	Institute of Geology, Vilnius	Lithuania
IGSB, Minsk	Byelorussia	RJ van de Graaff Lab, Utrecht	Netherlands
EIL, University of Waterloo	Canada	CIO Groningen	Netherlands
AECL, Chalk River	Canada	Rafter Lab, Institute of Geological Sciences	New Zealand
Geological Survey of Canada	Canada	University of Waikato	New Zealand

Continued

Table 6. Laboratories participating in FIRI (continued)

Laboratory name	Country	Laboratory name	Country
EHPL-Env, Ontario Hydro	Canada	Radiological Dating Laboratory, Trondheim	Norway
IOEE Chinese Academy of Sciences	China	Silesian Technical University, Gliwice	Poland
Rudjer Boskovic Institute	Croatia	A & E Museum, Lodz	Poland
Institut fur Physik, University of Aarhus	Denmark	ITN-Sacavem	Portugal
Institute of Geology, Tallinn	Estonia	Geological Institute, RAS	Russia
GSF, Espoo	Finland	Geographical Research, St. Petersburg State U.	Russia
University of Helsinki	Finland	Institute of Geography, RAS	Russia
IPSN/LMRE, Orsay	France	Institute of Ecology and Evolution, RAS	Russia
HIGL, Paris-Sud University	France	Institute of History of Material Culture, RAS	Russia
Tandetron-Gif	France	IQFR, Madrid	Spain
Universite Claude Bernard, Lyon	France	University of Granada	Spain
Umweltforschungszentrum Leipzig – Halle	Germany	Facultad de Quimica, Universitat de Barcelona	Spain
Leibniz, Universitat Kiel	Germany	Tandem Lab, University of Uppsala	Sweden
IUF, Universitat Koln	Germany	Universitat Bern	Switzerland
UFZ-CER, PRG, Halle	Germany	ETH, Zurich	Switzerland
Institut fur Bodenkunde, Universitat Hamburg	Germany	Department of Geology, NTU	Taiwan
Heidelberg University	Germany	Office of Atomic Energy for Peace	Thailand
DAI, Berlin	Germany	School of Geosciences, Queens University Belfast	UK
IGR, NLB, Hannover	Germany	Research Lab for Archaeology, Oxford	UK
Universitat Erlangen-Nurnberg	Germany	SUERC, East Kilbride	UK
LOIH, Insitiute of Physical Chemistry, Demokritos	Greece	NERC Radiocarbon Lab	UK
LOA, Institute of Materials Science, Demokritos	Greece	Lab of Radioecology, KIEV	Ukraine
Institute of Nuclear Research, HAS	Hungary	USGS, Reston	USA
Physical Research Lab, Earth Sciences Div., Ahmedabad	India	Beta Analytic Inc., Florida	USA
Physical Research Lab, Radiocarbon Dating Lab, Ahmedabad	India	NSF Arizona	USA
Birbal Sahni Institute, Lucknow	India	Geochron Labs, Cambridge, MA	USA
CRDIRT, JCPJ, Jakarta	Indonesia	CAMS/LLNL	USA
University College Dublin	Ireland	NOSAMS WHOI	USA
Kimmel Center, Weizmann Institute	Israel	INSTAAR, University of Colorado at Boulder	USA
RDL, University of Rome La Sapienza	Italy	UC Riverside	USA
		ISGS, Illinois	USA

That said, the FIRI results have emphasised that on average, ^{14}C laboratories (whether gas proportional, liquid scintillation or accelerator mass spectrometry) are providing accurate and precise results.

Conclusion

The results of the latest FIRI have demonstrated that there are no significant differences between the main measurement techniques (gas proportional counting, liquid scintillation counting and accelerator mass spectrometry) but there is evidence from some laboratories of small laboratory offsets relative to known age samples. There is also evidence in some cases for over- or under-estimation of measurement precision. Approximately 10% of all results were classified as extreme (outliers) and these results were generated by 14% of the laboratories. Notwithstanding good internal QA procedures, some problems still occur which can best be detected by participation in independent inter-comparisons such as FIRI where the results allow individual laboratories to assess their performance and to take remedial measures.

Acknowledgements

This work was supported by NERC (Grant Ref: GR9/03389) and the European Commission (Grant Ref: SMT4-CT98-2265). We also wish to express our gratitude to Mike Baillie, Marco Spurk, Roy Switsur, Glengoyne Distilleries, Ganna Zaitseva, Kh Arslanov, John Thomson and Alan Hogg who provided many of the samples.

References

- BRYANT C, I. CARMİ, G. COOK, S. GULLIKSEN, D. HARKNESS, J. HEINEMEIER, E. MCGEE, P. NAYSMITH, G. POSSNERT, M. SCOTT, J. VAN DER PLICHT & M. VAN STRYDONCK. 2002. Sample requirements and design of a inter-laboratory trial for Radiocarbon laboratories. *NIMB* 172: 355–359.
- GULLIKSEN S & E. M. SCOTT. 1995. TIRI report, *Radiocarbon* 37(2): 820–821.
- ISG, 1982. An inter-laboratory comparison of radiocarbon measurements in tree-rings. *Nature* 198: 619–623.
- LONG A & R. KALIN. 1990. A suggested quality assurance protocol for Radiocarbon dating laboratories. *Radiocarbon*, 32(3): 329–335.
- POLACH, H. 1989. ^{14}C CARE. *Radiocarbon*, 31(3): 422–431.
- ROZANSKI K, W. STICHLER, R. GONFIANTINI, E.M. SCOTT, R.P. BEUKENS, B. KROMER & J. VAN DER PLICHT. 1992. The IAEA ^{14}C intercomparison exercise 1990. *Radiocarbon* 34(3): 506–519.
- SCOTT E M, T.C. AITCHISON, D.D. HARKNESS, G.T. COOK & M. S. BAXTER. 1990. An overview of all three stages of the international radiocarbon intercomparison. *Radiocarbon* 32(3): 309–319.